

## **Microdata for Small Areas**

Mark Tranmer, Andrew Pickles, Ed Fieldhouse, Mark Elliot, Angela Dale,  
Mark Brown, University of Manchester  
David Martin, University of Southampton  
David Steel, University of Wollongong  
Chris Gardiner, Sheffield Hallam University

Address for contact:

Angela Dale, CCSR, University of Manchester, M13 9PL  
angela.dale@man.ac.uk

### **Acknowledgments**

We are grateful to the ESRC for funding this research under the 2001 Census Development Programme, grant number: H507255161. We are also grateful to many other colleagues for helpful comments at previous presentations of this work. We are very grateful to ONS for making data available to us. All data from the 1991 Census are Crown Copyright.

# Microdata for Small Areas

## 1. Introduction

The 1991 UK Census was the first from which samples of microdata (termed Samples of Anonymised Records or SARs) were released. Two files were extracted. The first was a 1 percent sample of households and all the individuals enumerated within those households. This file provides very great individual detail (eg 358 occupational categories) and allows linkage between all individuals within the same household. In recognition of this level of detail, the geographical threshold is set at about 1.6 million population – the size of the smallest standard region in 1991. The second file was a 2 percent sample of individuals with greater geographical detail - 120K population threshold – and a resultant reduction in individual detail; for example, there are only 73 occupational categories. Each file has been carefully designed to ensure the confidentiality of the respondents, based on the assumption that there is a three-dimensional trade-off between sampling fraction, individual detail and geographical detail.

Both files have been widely used within academia and, to a rather lesser extent, by policy analysts in central and local government. The topic coverage and detailed information contained in the SARs have allowed multivariate analysis at the appropriate unit of analysis - individual, family or household - with user-defined variable categorisations. By comparison with the pre-tabulated small area statistics the SARs have proved simple to use with standard software packages and have provided enormous flexibility.

The Individual SAR provides the greatest geographical detail available on any nationally-representative microdata dataset within the UK. The fact that the sample is drawn from a complete

population census means that, unlike traditional surveys, there is little additional cost incurred by increasing the sample size and using a geographically stratified sampling design. However, the geographical units identifiable in the Individual SAR have often been larger than desirable and are not always the most appropriate for the required analyses. This has led to user demands for finer geographical areas with the possibility of flexible aggregation to larger areas according to the analysis being conducted.

These requests, together with the recognition that, in terms of confidentiality, there is a trade-off between individual detail, sample size and geographical detail, has led to a proposal to establish the viability of samples of microdata for small areas - or Small Area Microdata (SAMs). The theoretical rationale for SAMs is set out below.

## **2. The rationale for small area microdata**

### *2.1 The role of place*

At the heart of the rationale for small area microdata is the debate over whether locality or geographical area plays an independent explanatory role over and above individual and household level characteristics. Do multivariate relationships between individuals and households vary between areas, or, does any apparent variation disappear once appropriate individual and household-level control variables have been introduced?

The disciplines of sociology, economics and geography both provide a considerable literature on the role of place and its influence on social processes. Within sociology, the importance of place

may be traced back to the community studies of the 1960s, followed, in the mid-1980s by research programmes such as the Economic and Social Research Council's Social Change and Economic Life Initiative, designed to examine the impact of local labour markets on economic activity and employment patterns. Within medical sociology, the local area has again been identified as playing a significant role over and above the characteristics of the individuals who live there. Macintyre (1997) distinguishes three kinds of spatial effects: *compositional* (the result of the characteristics of the people in the area); *collective* (the impact of group characteristics, for example, disruptive neighbours; or the impact of bright and able pupils on other pupils); and *context* (the physical environment and resources available). Thus high levels of unemployment in a particular area have a compositional element which results from the characteristics of the residents. A collective effect may be apparent through reduced social interaction between residents and a contextual effect evident in the deprived and run-down state of the area. This contextual effect may have a direct impact on employment opportunities through, for example, employer discrimination based on the address or postcode. It may also impact through the absence of facilities, such as public transport, that affect one's ability to get to work. In the example of health, local environment may have a direct impact through levels of pollution and noise, poor street lighting and lack of open spaces.

Within social geography, the role of place has been important for a considerable time. The time-geography of Hagerstrand (1973) explored how individuals moved through space in their daily lives. Research at the Centre for Urban and Regional Development Studies at the University of Newcastle has developed the potential for linkage between, for example, individual and spatial factors ( eg Coombes and Raybould, 1997). Other approaches have been aimed at demonstrating spatial variation, often through social atlases and mapping techniques, for example, Dorling's social atlas of Britain (Dorling, 1995).

From these examples it is possible to distinguish locality-specific research from the analysis of geographical variation. Locality specific research is primarily confined to a better understanding of specific places, often theoretically chosen to provide contrasts or driven by policy concerns.

Analysis of geographical variation, by contrast, usually has a national-level focus and seeks to establish and explain the role of place, usually using multivariate statistical modelling. Questions addressed by the latter approach might include:

- Is unemployment explained solely by individual characteristics?
  - what is the role of the local labour market and locality?
- Is the relationship between age, gender, social class and ill health the same across all geographical areas? Can we identify those areas that appear to be beneficial or detrimental to health? What are their salient characteristics?
- Do the housing/ employment experiences of minority ethnic groups vary by area, after controlling for individual & family characteristics? Can we identify the role played by area?
- What is the extent and pattern of multiple deprivation?

A specific example of how questions relating to the role of place can translate to policy research is provided by the Department of the Environment, Transport and the Regions (DETR ) in “Guidance on Local Housing Strategies” (DETR, 1998). This stressed that in the assessment of future housing needs by local authorities for their housing strategies it was:

‘ ... important to identify the differences between areas within the local authority when assessing needs e.g. rural versus urban, and between different types of people e.g. the needs of black and ethnic minorities may differ from other residents of the area’.

Empirical research which has addressed geographical variation has, to date, been conducted using census microdata, firstly through the Office for National Statistics' Longitudinal Study (Dale, 1993), and, since 1993, through the SARs. These datasets provide nationally representative samples with the necessary individual detail **and** geographical detail to identify geographical variation controlling for individual or household characteristics. Analysis has been facilitated through the availability of statistical models and software tools (eg MLWin, STATA) that are able to capture the multilevel structure of individuals, households and locality as well as the computing power that is needed to run the models. National-level census microdata has been used to identify area-level effects in women's employment (Ward and Dale, 1992); in health status (Gould and Jones, 1996); in unemployment (Fieldhouse and Gould, 1998) and in the development of local government policy indicators (Gardiner, 2000). However, these studies have been hampered by a geography that is not well tailored to the research question and is unable to identify the locale. Proposals for small area microdata are designed to overcome this constraint by providing microdata with sufficient geographical detail to support analyses of compositional versus contextual effects. In section 3 we provide examples of models developed in the context of unemployment, that demonstrate the value of being able to include locality as well as higher level geography.

## *2.2 An appropriately specified model*

Even where locality is not the prime focus it is important that statistical models recognise the geographically clustered nature of census data. If the natural clustering of these data is ignored in the models used, then there is a danger that the estimates of standard errors will be biased and this may lead to statistical significance (or non-significance) being falsely attributed to estimates. We

illustrate this by providing examples in section 3, below, of analyses with and without terms to represent geographical area.

### *2.3 As complementary to the Standard Tables and Census Area Statistics from the 2001 Census*

A third role for small area microdata would be as a complement to the aggregate statistics from the 2001 Census. The 2001 Standard Tables and Census Area Statistics will provide 100% data for a large set of pre-specified tables. Because of their fine geographical definition and the fact that they contain complete population data, the amount of detail in these tables is highly constrained.

Inevitably practitioners want further disaggregation and more complex tables in order to identify more precisely the characteristics of particular areas. Tables are typically constrained to include only 3 or 4 variables – for example, economic activity by ethnic group for the population aged 16-24 and 25-74 in England and Wales (Ethnic group Table ETH05). This table does not allow distinctions by sex or by detailed age-group – although economic activity levels vary considerable by both. Additional individual-level characteristics known to be important in explaining unemployment – for example, educational qualifications – cannot be included. Thus the policy analyst is not able to assess whether the variation in levels of unemployment between areas is explained solely by the characteristics of the residents. These questions have important policy implications.

Practitioners are often forced to bring together information from two or more discrete tabulations in order to establish the characteristics of an area. This runs the risk of making inferences about individual relationships from aggregate data. The problems of the ecological fallacy are well recognised (Robinson, 1950) but are hard to avoid when multi-way tables are not available for the

variables of interest. Small Area Microdata can be used to complement aggregate 100% census tables to overcome many of these problems. SAM can provide multi-way tables which avoid the need to base individual relationships on aggregate data. However, this benefit of microdata is only possible because SAM have other restrictions, necessary in order to ensure confidentiality. The most relevant restrictions in this context are a relatively small sampling fraction and a population threshold that is much higher than the most detailed geography available in the Census Area Statistics. The two data sources (SAM and CAS) are, therefore, complementary and section 3.4 discusses the methods by which this complementarity can be exploited.

### **3. Empirical examples of the value of SAMs**

#### *3.1 Data*

The Office for National Statistics has made available to us, under strictly controlled conditions, population data from the 1991 Census for seven local authorities<sup>1</sup>. Although the identities of these authorities are not known, they include inner city areas, rural areas and larger metropolitan cities. We extracted from this database a 10 percent sample of households (based on that used by ONS). These households have ED psuedo-identifiers which allowed further sampling to include geographical stratification whilst concealing the identity of areas.

From these data we have extracted samples of individuals at population thresholds of

---

<sup>1</sup> ONS supplied us, under contract, with anonymised and non-disclosive data from the 1991 Census for seven UK local authorities. The data were kept in a secure environment, with access limited to members of the project team. The data were returned to ONS when the work was completed. For the purposes of the contract, the Centre for Census and Survey Research was a supplier of services to the Registrar General for England and Wales and the 1991 Census Confidentiality Act applied.

approximately 5K, 10K, 15K and 30K, for sampling fractions of .05 and .10. These different sample designs have been assessed for analytical utility against confidentiality. Confidentiality assessment is discussed in section 4.

### *3.2 The role of place: Modelling unemployment with area distinctions*

In section 2.1 we outlined the theoretical reasons for wishing to establish the role of place. In this section we provide some empirical examples using the prototype samples of small area microdata described above.

Previous research on unemployment has demonstrated how the risk of unemployment varies at a range of geographical scales, even after controlling for individual characteristics. This might be, for example, at the level of the local labour market (Fieldhouse and Gould, 1998) or the level of the neighbourhood (Fieldhouse, 1999). Using an area classification on the 2% SAR, Fieldhouse and Tranmer (forthcoming) show that there is still considerable variation in unemployment between neighbourhood types after controlling for individual characteristics and labour market level differences. However, it is not currently possible to apply this knowledge to specific localities.

Small area microdata would overcome this constraint by providing identifiable local labour markets and, depending on the population threshold, might also provide identifiable localities. However, in the analyses presented here confidentiality requirements prevent us from identifying the geographical areas used. Nonetheless, in the context of an analysis of unemployment, we can show the proportion of variance that relates to the area of residence rather than to individual

characteristics.

Figure 1, based on 1991 census test data for seven local authorities, shows a decomposition of variance between the individual, SAM areas of 30K population, and the local authority district. This analysis of unemployment includes a full set of individual characteristics: ethnic group, UK born, higher qualifications, age group and family composition. Figure 1 shows there is a substantial amount of variation between sub district areas, even with a fairly crude (30K) geography. This is above and beyond any variation between districts. Using this type of approach it is possible to understand the extent and correlates of unemployment at a variety of different scales. Furthermore, by examining the predicted and residual values it is possible to see which areas have higher than expected levels of unemployment given their social and economic composition.

For the null model we see that around two-thirds of the area-level variation – that is, the variation at areas above the individual level – is at the 30K population threshold. When we allow for a range of socio-economic characteristics in the full model, there is less area level variation as the composition of the area is now included in the model. However, the majority of the area level variation remains at the 30K threshold level rather than at district level.

**Figure 1 about here**

Figure 2 shows a decomposition of variance in unemployment using a hierarchical (nested) multilevel model within a single local authority, which we call ‘Ambrosia’, with three different levels of geography defined. The explanatory variables are the same as in the previous model. Precedence is given to higher levels; the extent of variation at the smaller scale, in both the null

model and the model with controls, shows that for this type of analysis, designed to identify spatial differences in unemployment, it is important to have the finer geography. This is likely to apply to almost any analysis, as smaller areas are always more homogenous than larger ones. The identification of area-level differences tells us that there is an explanatory role for place, even when a full set of individual control variables are included.

**Figure 2 about here**

Figure 2 is also important in telling us which geography might be the most useful. When modelling the four different levels simultaneously there is substantially more variation at the lower level of geography (7.5k threshold) than there is at the higher (30K threshold). This would suggest that, for the analysis of unemployment at least, it is important to have fine geographical detail in small area microdata.

We pointed out, above, that this analysis is using confidential test data and therefore we cannot identify the areas from which the data are drawn. In the 'real-world' we would wish to use the smallest available area as the locality and then to aggregate localities to form travel to work areas. Used at a national or regional level, we could extract residuals to identify those areas which have higher unemployment even after including a wide range of individual-level characteristics. We might also include explanatory variables at each of the area-levels identified in order to assess the reasons for the area-level effect.

Sample size for SAM areas is an important consideration, but should not be approached purely

from the basis of the size required to achieve a given level of precision for individual areas taken in isolation. As discussed in section 3.4, multilevel modelling provides one way of improving local area estimates via shrinkage by considering them in the context of the other areas. However, analysts often want to extract residuals and these are more heavily influenced by small numbers. It is therefore a priority to maximise sample size to the extent that this is compatible with ensuring confidentiality and sufficient individual detail.

### *3.3 An appropriately specified model*

Normal regression techniques do not take into account the clustered nature of individuals with particular characteristics. This can affect the parameter estimates and their sampling errors, with the result that characteristics may appear significant when, with a more appropriately-specified model, they are not.

Table 1 shows the change in coefficients for various ethnic groups within a full model of unemployment. In column 1, using a logistic model, geographical area is not included. In columns 2-4 different random intercepts for different sized geographical areas are included in a series of multilevel models. The final column shows the effect of ethnic group when geography is added as a dummy variable in a logistic regression.

Without area included, we can see that both Indian and Pakistani groups are significantly more likely to be unemployed than the reference category (whites). However, once area is included -

even as a dummy variable - Indians are no longer significant in the model. The addition of an area-level indicator to the model reduces the standard error of the estimate, although only slightly, and considerably reduces the size of the parameter estimates. The benefit of the multilevel model over the logistic model is that it allows explanatory variables to be added at the area level and, as we see below, also allows the use of shrunken estimates to improve the precision of local area estimates.

The results from this table are also of interest because they identify unemployment patterns that are counter to those usually found at a national level. For example, by comparison with the white reference group, unemployment is higher for Pakistanis than for Bangladeshis. This demonstrates the importance of area-level differences. (A re-analysis using a 10% sample produced very similar results.)

**Table 1 about here**

### *3.4 An addition to the Census Area Statistics*

We argued in section 2.3 that microdata can provide important cross-tabulations and disaggregations not available with the 100 per cent data. For example, Simpson (2000) has identified the need for detailed economic activity rates by age and sex for each ethnic group in order to make labour force forecasts in multi-ethnic areas. However, the small area sample sizes typical of SAM means that local area estimates from such tabulations have low precision and therefore large confidence intervals. To compensate for small sample sizes, synthetic estimation

can be used to combine the SAM with 100% aggregate data from the census for the same geographical area. Synthetic estimation ‘borrows’ information from relevant other sources to improve point estimates and their precision.

Here we have used a number of different models to demonstrate the potential of synthetic estimation – both to improve the precision of estimates from small area microdata and also to extend 100% aggregate tables from the census.

*i. Using population data from the census tabular output to improve SAM estimates*

An obvious source of local population data are the aggregate tables from the census. The margins on the SAM tabulations can be compared with tabulations available from the 100% census tables, recoding the SAM variables, if needed, to achieve comparability. As a consequence of the sampling variance in the SAM, the SAM tabulation will not exactly match the 100% table, and sometimes, particularly for small areas, the discrepancy can be large. Various equivalent approaches to achieving consistency are possible.

Simpson (Bradford City Council, 1996) used Iterative Proportional Fitting to ensure that local area disaggregate estimates of economic activity matched aggregate totals. In this algorithm, described in detail by Rees (1994), the frequencies in the disaggregate tabulation are successively scaled to match the known margins, while maintaining the interactions of the original microdata. Rindskopf (1992), approaching the problem as one of missing data, provides an alternative algorithm that uses composite link functions in a generalized linear modelling framework, one that can be more easily implemented in standard software.

Viewing the problem as one of missing data suggests the approach of using the sample SAM to fill in the missing data in the 100% table, in other words to generate *population* microdata. This has been long used in geography under the name of micro-simulation and in this specific context by Williamson et al (1998). Recent developments in statistics in the field of multiple imputation (e.g. Rubin, 1987; Schafer, 1995) and other Bayesian based simulation estimation provide additional theoretical support together with the prospect of the integration of these methods into standard software (e.g. STATA's HOTDECK procedure, Mander and Clayton, 1999).

ii. *Using SAM tabulations from other areas to improve estimates for a given area*

SAM based local area estimates vary one from another as a consequence of both true local variation and sampling variance. It has long been known (James and Stein, 1961) that improved estimates of individual values in a sample that are subject to measurement error can be obtained by shrinking the naïve values towards the sample mean. Longford (1999) sets out how this can be done using small area microdata. In a number of cases, even where the estimation technology looks sophisticated, the synthetic estimate boils down to a weighted average of the naïve local area estimate and the overall mean. The weight depends upon the variance among the local areas, the sample size available in the local area and that available in the whole area of interest. In the simple case the 'shrunk' estimate of a local area proportion is

$$p_i^* = (\sigma^2 p_i + v_i p) / (\sigma^2 + v_i)$$

where  $p_i$  and  $p$  are the naïve local and overall estimated proportions,  $\sigma^2$  is the between area variance in the true local area proportions and  $v_i$  is the local area sampling variance. Compared to the naïve

estimator the mean square error for this shrunken estimator is expected to be smaller by a factor of  $1/(1 + v_1/\sigma^2)$ . In Longford's example of economic activity rates among young men the estimated standard errors of the local area estimates were 1.13 to 2.15 times smaller than the naïve estimates, the greatest benefits being obtained for the areas with the smallest sample size.

In the univariate case these methods can be implemented in a spreadsheet, but more generally are not difficult in any software implementation of random effects models that allow the estimation of random effects/shrunken residuals or empirical Bayes estimates. These include MLWiN (Goldstein et al, 1998) and the STATA procedure GLLAMM (Rabe-Hesketh, Pickles & Taylor, 2000).

iii. *Using SAM tabulations for some other correlated variable*

One can borrow information about spatial variation from groups other than the particular one that is the focus of study by the use of multivariate shrinkage. For example, Longford (1999) illustrates how information on economic activity rates among whites can be used to increase the precision of estimates for ethnic minorities. Imprecise estimates of spatial variation among non-whites are here shrunken towards the much better estimated pattern of variation among whites. The improvement in precision so derived can be considerable.

iv. *Using SAM and population data for a correlated variable*

Standard survey analysis provides numerous examples of the potential benefits in precision available from the use of 'regression estimators' and 'post-stratification', by variables that characterise the population structure. Tranmer and Steel (1998) found that housing tenure, housing

type and older age groups were important variables to characterise population structure at local area level, and referred to these as 'grouping variables'. If, for example, local housing tenure totals are known, and housing tenure is related to unemployment, then more precise estimates of local unemployment can be obtained by an appropriately weighted sum of the tenure-specific unemployment rates (where the weights for each area are the known tenure proportions) than by using the sample mean by itself. The proposed approach essentially involves a combination of the ideas from (i) and (iii) above and could well offer an approach of considerable power, although it is still in its infancy.

In the following section we use an empirical example of how small area microdata can be complemented by using standard census tabular output for the same variables and the same areas and also by using SAM tabulations from other areas. These methods can be used to improve the precision of sample estimates for small areas and can also be used to extend 100% census area tabulations beyond those already available.

### *3.5 An applied example using Small Area Microdata with (i) SAM tabulations from other areas and (ii) 100% aggregate census tables*

The examples reported below are designed to:

- a) improve the precision of estimates based on the microdata and
- b) add extra dimensions to 100% Standard Tables or Census Area Statistics by synthetic estimation.

Both can be achieved by combining microdata estimates with aggregate data. Appendix 1 sets out a simple schematic table to demonstrate this. A series of modelling approaches have been used, all based on the example of predicting the unemployment rate for non-whites in one unknown local authority, termed Ambrosia. The data used are prototype SAM representing a 5% sample for 30 sub-areas of Ambrosia, each with a 15K population threshold. Each approach is described in turn, starting with the most simple. The results from applying each model are shown in Table 2.

*Approach 1: Using small area microdata and simple logistic regression (M1)*

A simple logistic regression model may be used to predict the unemployment rates for non whites in each of the 30 areas, based on fitting the model:

$$\text{Logit}(p_{ij}) = \mathbf{b}_1 + \mathbf{b}_2 + \Lambda + \mathbf{b}_{30}$$

to the non whites data only, where 30 dummy variables (and no intercept) are included, one for each area. The estimates are the logits of the estimated unemployment rates for each area and can be easily transformed into predicted probabilities. Confidence intervals may also be derived from the standard errors of these estimates. Although this method uses all the SAM data, the predicted unemployment rate for each area is only based on the SAM data for that area.

*Approach 2: Using small area microdata in a multilevel logistic framework. (M2)*

$$P_{ij} = p_{ij} + e_{ij}$$

In this approach we use data for all 30 areas and both whites and non-whites for the estimation of unemployment rates among non-whites for a particular SAM area. The model may be fitted in any multilevel modelling package and we used MlwiN, based on the penalised quasi likelihood (PQL) procedure. The model is

Where

$$\text{Logit}(p_{ij}) = \mathbf{b}_0 + \mathbf{b}_1 x_{ij} + u_{0j} + u_{1j} x_{ij}$$

$x_{ij}$  is a variable that takes the value 0 for white and 1 for non-white. Thus we can obtain estimates for each area using

$$\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1, \hat{u}_{0j}, \hat{u}_{1j}$$

where

$$\hat{\mathbf{b}}_0, \hat{\mathbf{b}}_1$$

are the estimated coefficients of the constant and non-white respectively, and

$$\hat{u}_{0j}, \hat{u}_{1j}$$

are the area level residuals of these coefficients.

Estimates may be obtained for each area on the logit scale with confidence bounds from the corresponding standard errors obtained from the fixed and random parts of the model. These can then be transformed to the probability scale. This approach uses all areas in the analysis and 'borrows strength' from the other areas to derive the estimates for each area.

*Approach 3: Combining small area microdata with marginal information from 100% tabular data in a multilevel model (M3)*

In this model, information about some relevant margin obtained from the 100% aggregate tables for each SAM area is included as an explanatory variable.

$$\text{Logit} ( p_{ij} ) = \mathbf{b}_0 + \mathbf{b}_1 x_{ij} + \mathbf{b}_2 \bar{Y}_j + u_{0j}$$

Where

$$\bar{Y}_j$$

is the proportion of all people in the population of interest who are unemployed in area j. This

information is based on the appropriate cells for the margins of the table (see Appendix 1). Because we are using a multilevel approach we are also 'borrowing strength' from the other areas.

*Approach 4: Combining small area microdata with marginal information from 100% tabular data using a logistic model (M4)*

We can simplify this approach by using a logistic model so that we are still combining data from 100% tabular data and SAM, but we are no longer borrowing strength from other areas as we were in the multilevel approach. The logistic model is very easy to fit in packages such as SPSS. The logistic model is as in (3) above but without the area level residual terms and is written as:

$$\text{Logit}(p_{ij}) = \mathbf{b}_0 + \mathbf{b}_1 x_{ij} + \mathbf{b}_2 \bar{Y}_j$$

*Approach 5: Combining small area microdata with information from two margins of a 100% table in a multilevel model (M5)*

This final multilevel approach uses the 100% aggregate data for both the unemployment and non white margins of the table, together with the sample statistics from the SAM. The model is

$$\text{Logit}(p_{ij}) = \mathbf{b}_0 + \mathbf{b}_1 x_{ij} + \mathbf{b}_2 \bar{Y}_j + \mathbf{b}_3 \bar{X}_j + u_{0j}$$

Where

$$\bar{X}_j$$

is the proportion of non whites in the SAM area, derived from the 100% tabulations.

Using each model we have estimated the unemployment rates for non-whites in one local authority district (Ambrosia) for five of the 30 areas defined in the microdata files. These are shown in table 2 with their associated confidence intervals. We have also obtained the true values of the

unemployment rates (shown as 'truth' in the table) from the corresponding 100% data from which we drew the sample. The population is the economically active.

Areas 1,2 and 3 have fewer than 100 non-whites in the total population of the economically active and 2, 3 and 4, respectively in the sample of microdata. Areas 4 and 5 have much bigger populations of economically active non-whites with samples of 57 and 60, respectively, in the microdata.

From table 2 we can see that the estimated rates of unemployment are much improved by borrowing strength from other areas or by including the marginal values of unemployment from the 100% aggregate tables as a predictor. This is particularly important in areas with very small numbers of economically active non-whites in the sample. In general, the logistic model (M4) is almost as good as the multilevel model (M3) but much simpler to apply. Interestingly the addition of the non-white marginal to model 5 does not lead to an obvious improvement in the estimates. In general these results show that imprecise estimates with large confidence intervals that arise from the SAM can be greatly reduced by combining the SAM data with other census data under a model based approach. Prediction intervals could also be calculated and these would behave in a similar way to the confidence intervals. The methods illustrated can be used to provide additional details in aggregate census tables. The results indicate the importance of specifying SAM that is compatible with census geography for other census data sets.

Although these results need to be extended by producing models related to different topics and using a broader range of geographies, they indicate very promising way forwards by which different outputs from the census can be used to complement each other.

### *3.6 Small Area Microdata as a pre-cursor to using an on-line tabulation service*

If an on-line tabulation system is provided by ONS which allows access to 100 percent data at a speed and cost acceptable to the user this will be preferable to basing tables on sample data.

However, tables released by ONS will, of necessity, be carefully monitored for confidentiality and may not provide all the multivariate break-downs required. Even where required break-downs are allowed, SAM can provide a test-bed to ensure that tables requested are, indeed, those needed by the analyst. In addition, where complex sets of derived variables are needed it will be important to test these fully before submitting to ONS. In all these circumstances small area microdata can play a valuable role.

## **4. The confidentiality risk associated with different population thresholds and different sample sizes**

Confidentiality is a crucial issue in the above proposals and has been extensively investigated. In this section we outline the framework used to address confidentiality and then report results which assess the risk to confidentiality of different sample sizes with differing population thresholds.

The decision by ONS to release the 1991 SARs was taken in the light of extensive evidence on the risk of disclosing information about individuals or households through identification. An ESRC working party chaired by the late Cathie Marsh estimated that the per record risk of identification of individuals in the SARs was negligible (Marsh et al, 1991). More recently, we have been able to reassess this work with the benefit of access to a much wider range of data sources. This has

confirmed the level of risk estimated by Marsh et al and suggested that, if anything, the risk is lower (Dale and Elliot, 2001; (<http://les1.man.ac.uk/ccsr>). The results of this work therefore allow us to use the 1991 SARs as a base-line against which to assess the risk of identification from a range of specifications for small area microdata.

Further work on confidentiality (Elliot et al, 1998; Elliot and Dale, 1999) has provided a broader framework and more extensive range of measures for assessing the risk of identification. This included identifying the scenarios under which attempts at disclosure are most likely to occur and the kinds of local information that may be used in such an attempt. This provides the basis upon which standard 'key variables' are defined - that is, those variables most likely to allow identification of an individual in the sample dataset by matching with an outside dataset. This work has also shown that there is a clear trade-off between geographical detail, sample size and detail relating to individuals and households (Elliot and Dale, 1998). Table 3, using the local authority population data described above, demonstrates this using a measure of disclosure risk - the percentage of sample uniques that are unique in the population – which provides a more realistic assessment than the level of uniques in either the sample or the population. We are able to calculate this measure for a file specification comparable to that of the 1991 Individual SAR and then assess alternative specifications against that base-line. Seven key variables are used: age, sex, ethnic group, migration status, economic status, tenure and marital status. These variables were selected as holding information that could be obtained about an individual by informal methods, for example talking to neighbours and personal knowledge.

Based on the parameters of the 1991 Individual SAR file, and the variable coding used with that file, about 17% of sample uniques on this set of key variables are also population unique. This

figure provides a benchmark against which to assess alternative specifications. To test the extent to which reduced individual detail offsets the effect of increased sample size and reduced population threshold we have grouped age into 5-year bands (19 categories); reduced the number of ethnic categories to four and reduced economic status from 11 to 4 categories. We can see that, with this reduced detail, we get similar values to that for the SAR specification (17.5%) at very small geographies if we keep the sampling fraction low, or, at larger geographies (8K or 16K) if we increase the sampling fraction to 5%. Generally, these figures suggest that sampling fraction has a much larger impact on the risk of correct identification than geography.

### **Table 3 about here**

An alternative basis for estimating risk has been developed by Elliot (1998) and formalised by Skinner. A detailed description of the method, termed Disclosure Intrusion Simulation (DIS) is given in Appendix 2. It has the benefit that it does not require population data. Extensive validation shows that, within a very wide range of parameters, it generates the same results, based solely on a sample, as are generated using population data (Elliot, 2000). From this general method, two specific measures have been derived:

- i. *The probability of a correct match given a unique match.* This estimates the probability of a record from an external file correctly matching a record in a sample given that it has matched that record uniquely (assuming the data in each file are recorded identically, using the same coding regime, for the same time period). It represents the probability faced by an intruder who has matched a sample file with an external file and identified all the unique matches.

- ii. *The probability of a correct unique match.* The probability that an arbitrary record from an external file will correctly match with a unique record in the target file. This represents the probability an intruder would face trying to match any given record in an identification file to a target file.

These measures have been used to compare various formulations of SAM files against a baseline measure of the 1991 Individual SARs. Two types of analysis have been conducted:

*Additional impact analyses.* These take a standard key variable set consisting of age, sex and marital status and examine the effect of adding each other variable to that basic set.

*Scenario based analyses.* This uses a set of disclosure risk scenarios defined by Elliot and Dale (1998) to identify a set of key variables selected to represent the most likely scenarios in which an attempt to breach confidentiality might occur. The scenarios were:

*Scenario la: Demonstrative Political attack*

Variables: age, sex, marital status, economic activity, ethnic group, country of birth, migration in the last year, tenure, and long term limiting illness.

This assumes a political group acting in collusion with respondents who provide them with copies of the information given on their census forms. We have avoided variables which give information about other individuals, apart than the colluding agent, on the assumption that this would go against the underlying rationale for the attack.

*Scenario lb: Demonstrative Political attack full set.* This is the same as above except that all.

variables are used. This can therefore be used as a measure of whole file risk and allows comparison between using a restricted set of keys and all the variables in the file.

*Scenario 2a: Private Database Cross Match:*

age, sex, marital status, number of cars, number of dependent children, workplace, distance of journey to work, number of residents.

The variables in this scenario are based on those most likely to be available to a large commercial company collecting lifestyle data.

*Scenario 2b:* As with 2(a), except with the addition of: occupation employment status, number of earners, tenure.

*Scenario 3: Journalist using information about colleagues:* sex, marital status, economic activity, occupation, industry, age, address one year ago, ethnic group, long term limiting illness, presence bath and wc at home.

Assessments were conducted on sampling fractions of 5% and 10% with geographical thresholds for 5K, 10K, 15K and 30K for a range of variable specifications. Those reported here represent a 5% sampling fraction with 5K and 10K population thresholds. They were run on data for two of the LADs supplied by ONS. The variable specification for the files is shown in Appendix 3, along with the specification for the 1991 Individual SAR, which provides the base for comparison.

*4.1 Results*

The additional impact analyses are shown here only as summary values in table 8, discussed below.

Results from the various scenarios are shown in tables 4-7 for the two methods of assessing risk and using data from both LADs. Thus Table 4 shows the probability of a correct match given a unique match under the various disclosure risk scenarios for Ambrosia, with figures for the 1991 SARs and two possible SAM files — with 5K and 10K thresholds respectively.

The probability of a correct match given a unique match and the probability of a correct unique match are both measures of the likelihood of an intruder succeeding in identifying an individual in an anonymised dataset. Given this, it is important to stress the low values in all cells in tables 4-7. The 1991 SARs have been demonstrated to be effectively safe using other measures (Dale and Elliot, 2001) and our concern here is to compare values for the two SAM files (5K and 10K population thresholds) with the 1991 Individual SAR for the appropriate SAR area. In table 4 we can see that, under all the scenarios, SAM1 produces lower probabilities, closer to those for the 1991 SAR, than SAM2. Similarly, in table 5, showing the probability of a unique correct match, SAM1, with the larger population threshold, consistently shows the lower risk level.

In general, the second LAD, (termed Carnation) (tables 6 and 7) shows a higher level of risk than Ambrosia for both the SAM and the SARs. This reflects the different population structure of the two areas and indicates the extent to which there is geographical variation in risk levels. However, the relative risk values of the SARs and the two SAM files remain similar.

Table 8 summarises all the analyses conducted for both LA areas and the 1991 SAR specification. We can conclude that both SAM files show risk probabilities of the same order as the SARs,

although that for SAM2 (5K threshold) is consistently larger than that for SAM1 (10K threshold). This suggests that, manipulating the three main risk parameters (population threshold, sampling fraction and individual detail), it is possible to specify small area microdata that have levels of disclosure risk very similar to that of the 1991 Individual SAR.

**Tables 4, 5, 6, 7, 8 about here**

## **5 Specifying the SAMs**

Having established the analytic value of small area microdata and presented a thorough assessment of the disclosure risk of such samples, we now move on to discuss the proposed basis for building SAMs.

### *5.1 What sampling fraction?*

Whilst a larger sample size is desirable, risk measurements, discussed above, show that an increase from a 5% to 10% sample would have very considerable implications for confidentiality and the resultant loss of detail is unlikely to be an acceptable trade-off. We have therefore opted for a 5% sample as the smallest compatible with useful data.

### *5.2 What geography?*

There are four key aspects to the question of defining a SAM geography. The first concerns the

population threshold which SAM areas must exceed to meet confidentiality requirements. This has implications for the second aspect, the choice of building blocks used to construct individual SAM areas. Essentially this is a choice between wards and census output areas. Third, there must be a set of agreed design criteria by which individual building blocks (output areas or wards) will be aggregated into SAM areas. Finally there must be a method to implement that design. Each is discussed in turn and all have been the subject of extensive user consultation.

### *5.2.1 What is the appropriate threshold size?*

Geographical thresholds from 5K to 30K were considered. It was concluded that a threshold at the lower end of the range would maximise flexibility and maximise the range of policy and administrative areas that could be derived from SAMs. Therefore the final tests for confidentiality reported above, were based on areas from 5K to 10K. Much greater geographical flexibility would be achieved with a 5K threshold, although the risk to confidentiality is slightly higher than with a 10K threshold.

### *5.2.2 Choosing the building blocks for a SAM geography.*

The new Census Output Areas (OAs) (Martin, 1998a) – which replace EDs as the lowest level of census output - and wards were both considered as the building blocks for a new SAM geography. There are a number of advantages and disadvantages associated with each. It is important to recognise that identification of these advantages and disadvantages is not independent of the temporal dimension and will reflect changes in cultural, political and policy concerns. One objective, therefore, has been to try to identify possible future trends in the definition and

identification of spatial areas for research and policy analysis. Two possibly conflicting patterns can be seen at present. On one hand a series of government initiatives have led to the creation of a series of action zones, which do not necessarily fit into existing (or likely future) census output areas. In contrast there have also recently been government proposals for new, or revised, data and information which do relate to more traditional concepts of spatial area, such as wards and EDs. The following discussion needs to be considered, therefore, within this framework.

Census OAs have the obvious advantage of their small population size (likely to average around 100 households). This will ensure flexibility in zone design and maximise the potential to derive a SAM geography of uniform size and shape. Moreover with OAs derived from postcodes, they have the advantage of facilitating matching to a wide range of data sources. Finally, OAs will aggregate relatively easily into higher levels of geography including census wards, districts and other statutory areas.

On the downside, OAs will not be designed until census enumeration has been completed. This may impose a serious delay to the development of SAM geography, since the required OA boundaries may not be available until the Census Area Statistics are published. Also, the small size of OAs relative to the larger SAM area sizes being considered (between 15,000 and 30,000) would impose a relatively heavy computational demand (in total there are expected to be around 200,000 OAs). Finally, depending on the design criteria used, aggregations of OAs within local authority districts may result in SAM areas which do not equate, nor aggregate neatly, to any other geographical units, including wards.

The main advantage of using wards is that they have inherent meaning among a wide range of

census users, notably within local authorities, and are used for a wide variety of purposes. The relevance of this level of geography for the SAMs has been reinforced by recent government announcements about policy and resource allocation mechanisms. For example, publication of the Indices of Deprivation (DETR, 2000) represents an important move towards the presentation of information at ward level on a consistent basis across the whole country. On 10th October 2000 the Government announced that the Neighbourhood Renewal Fund would distribute £800 million and that eligibility and basis of distribution of this fund would be the Indices of Deprivation 2000. In a closely related initiative the Social Exclusion Unit (2000), in a report entitled “National Strategy for Neighbourhood Renewal: a framework for consultation” proposed the creation of a set of Neighbourhood Statistics. The initial version of these would be a national ward-level dataset. While the population threshold on SAM areas is likely to require that many wards be grouped together, larger wards will be retained as SAM areas in their own right. In contrast to OAs, the ward boundaries at the time of the census will be known in good time for SAM area creation, and the much smaller number of wards (10,000) compared to OAs will enormously reduce the computational burden involved. Finally, a SAM geography built from wards would have the important benefit of easy linkage to the standard tables which will be published at ward level.

The main disadvantage of using wards is that ward boundaries are subject to continual change that may, in some cases, be substantial. They also show considerable variation in their population size and shape, particularly between different counties and between urban and rural environments. It is perhaps worth noting, however, that in mitigation the Social Exclusion Unit concluded that although neighbourhoods straddle ward boundaries and that wards vary a lot in size nevertheless wards are the best boundaries that exist [for tracking outcomes in small areas].

Although it is potentially a good feature that some wards will be retained as SAM areas, large numbers of wards close to or just below the SAM population thresholds will significantly reduce the available range of workable SAM area configurations, resulting in wide ranges in SAM area population. Combining two wards with just below-threshold populations will result in a SAM area with almost twice the threshold population. This is not a problem with OAs due to their small population sizes relative to the SAM areas. Finally, the large size of wards, and their potential to include areas of diverse social and economic characteristics sets limits on the effectiveness of imposing social homogeneity as a main design criteria in defining SAM areas.

On balance, and after considerable consultation with users, we recommend that wards should be used for the SAM geography. However, to overcome the worst problems of variation in size we propose that wards with populations at least double the SAM threshold (7-10K) – which may be termed ‘superwards’ - should be split into separate SAM areas. Only about 3% of wards in England and Wales are above 15K in size so this operation would only be required for a fairly small number. This subdivision would reduce the variance in the mean SAM population, thereby producing more homogeneous SAM areas. More, smaller SAM areas would also give greater flexibility in building to higher level geography and in multilevel modelling applications. However, splitting ‘superwards’ will incur additional computational requirements and would also require OA boundary data, adding both cost and time to the final output.

*5.2.3 Design criteria by which individual building blocks (output areas or wards) will be aggregated into SAM areas*

Apart from size, the main criterion to be applied would be homogeneity. Homogeneity of areas is particularly important when undertaking geographical analysis. For example, in an analysis of individual or household deprivation by area, it would be important for those areas to be as homogenous as possible with respect to deprivation indicators. If aggregated areas are very heterogeneous then, by definition, geographical variations and concentrations become diluted.

If wards are used as building blocks then SAM boundaries would naturally fall within the boundaries for many other administrative areas – for example, local authorities and most health authorities. Where wards need to be combined to meet the required threshold areas it will be necessary to decide whether these should be contiguous where continuity might decrease homogeneity. These decisions will be subject to user consultation - that may also raise the need for groups of wards to reflect policy requirements - for example, eligibility for European structural funds.

#### *5.2.4 Implementation methodology*

The agreed criteria would then be used to develop zone design algorithms. The main objective is to develop algorithms to produce areas which best comply with user and confidentiality requirements. We suggest a two stage strategy whereby the simple ward SAM are produced using automated zone design procedures (AZP) (Openshaw and Rao, 1995; Martin, 1998a, 1998b) to group sub-threshold wards. The SAM file could thus be delivered in the first wave of 2001 census output. Stage 1 will be relatively straightforward in terms of assessing all wards within each local district and aggregating to meet the population threshold. In the second stage, AZP would be targeted on the ‘superwards’, identified during stage 1, for the assembly of SAM areas from OAs. This would not require a 2-stage release of SAM but would represent the supply, by ONS, of an additional variable

for the SAM dataset.

## **6. Conclusions**

Small Area Microdata have the potential to provide a unique source of geographical and individual-level information. With the objective of linkage to the 2001 Census standard tables in mind, and working within the constraints of confidentiality and geography, we have produced a suggested level of disaggregation of the variables to be included in a SAM file. This specification (a 5% sampling fraction, a population threshold between 5-10K and a ward-based geography) is shown in Appendix 3. It is designed to allow relevant localities and areas to be identified whilst still retaining a large enough sample and sufficient individual detail to support multivariate analysis. The confidentiality risk from such a file is assessed as broadly comparable to the risk of the 1991 Individual SAR.

Preliminary analyses have demonstrated the value of being able to include appropriate geographies in multivariate models. We have also shown how synthetic estimation can enhance the value to both microdata and aggregate statistics. Although further work is necessary on the exact specification of SAM files, there is every indication that they could form an exciting addition to the range of outputs from the 2001 Census.

## **References**

Bradford City Council (1996) *Forecasts of the labour force*: technical report. Corporate Services, City Hall, Bradford BD1 1HY.

Coombes, M and Raybould, S (1997), *Modelling the influence of individual and spatial factors*

underlying variations in the levels of secondary school examination results, *Environment and Planning A*, 29: 641-658.

A. Dale (1993) "The Potential of the OPCS Longitudinal Study for urban and area-based research," *Environment and Planning A* 25, 10, 387-1398.

Dale, A, and Elliot, M. J. (2001) Proposals for the 2001 SARs: an assessment of disclosure risk. *Journal of the Royal Statistical Society (Series A)*.

Department of the Environment , Transport and the Regions (1998), Departmental Guidance on Local Housing Strategies, May 1998;

Department of the Environment , Transport and the Regions (2000) *Indices of Deprivation*, Regeneration Research Summary, Number 31.

Dorling, D. (1995) *A new social atlas of Britain*, (Chichester:Wiley)

Elliot, M.J. (1998) DIS: Data Intrusion Simulation - a Method of Estimating the Worst Case Disclosure Risk for a Microdata File. Proceedings of 1st International Symposium on Linked Employee-Employer Records, Washington; May 1998.

Elliot, M.J. (2000) "DIS: A New Approach to the Measurement of Statistical Disclosure Risk", *Risk Management: An International Journal* 2 (4).

Elliot, M. J., and Dale, A. (1998) Disclosure Risk for Microdata. Report to the European Union ESP/ 204 62/DG III.

Elliot, M. J. and Dale, A. (1999) Scenarios of Attack: The data intruder's perspective on statistical disclosure risk. Invited paper for special edition of *Netherlands Official Statistics*. Spring 1999.

Elliot, M. J., Skinner, C. J, and Dale, A. (1998) Special Uniques, Random Uniques and Sticky Populations: Some Counterintuitive Effects of Geographical Detail on Disclosure Risk. *Research in Official Statistics*; 1(2)

Fieldhouse, E A; Gould M I; 1998. Ethnic Minority Unemployment and Local Labour Market Conditions in Great Britain. *Environment and Planning A* vol. 30 no. 5 833-853.

Fieldhouse, E A; 1999. Ethnic minority unemployment and spatial mismatch: the case of London. *Urban Studies* Vol. 36 no. 10, pp 1569-1596

Fieldhouse, E and Tranmer, M (forthcoming). Concentration effects, spatial mismatch or neighbourhood selection? Exploring labour market and neighbourhood variations in male unemployment risk using census microdata from Great Britain.

Gardiner (2000, forthcoming), "Proposals towards the use of Samples of Anonymised Records and Customised Output from the UK 2001 Census of Population for development of local government

indicators for policy and resource allocation”, *Local Government Studies*

Goldstein, H (1995) *Multilevel statistical models*, 2nd Edition. Arnold: London

Gould MI and Jones K (1996) Analysing perceived limiting long-term illness using UK Census microdata *Social Science and Medicine* 42, 857-869

Hagerstrand, T (1973) The domain of human geography. In Chorley R J. ed. *Directions in geography*. London: Methuen.

James, W. & Stein, C. (1961) Estimation with quadratic loss. In Proc. 4th Berkeley Symp. *Mathematical Statistics and Probability*, vol 1 pp.361-379. Berkely: University of California Press.

Longford, N.T. (1999) Multivariate shrinkage estimation of small area means and proportions. *J. Roy. Statist. Soc.* 162A 227-246.

Macintyre, S. (1997) ‘What are spatial effects and how can we measure them?’ in *Exploiting national survey and census data: the role of locality and spatial effects*, pp.1-18 in A.Dale, (ed) CCSR Occasional Paper No.12, CCSR: Manchester

Mander, A. & Clayton,D. (1999) Hotdeck imputation.sgl 16, *Stata Technical Bulletin* 51.

Marsh, C.; Skinner, C.; Arber, S.; Penhale, P.; Openshaw, S.; Hobcraft, J.; Lievesley, D.; Walford, N. (1991). The Case for a Sample of Anonymized Records from the 1991 Census. *Journal of the Royal Statistical Society Series A*,154, 305-340.

Martin, D. (1998a) 2001 Census output areas: from concept to prototype *Population Trends* 94, 19-24

Martin, D. (1998b) Optimizing census geographies: the separation of collection and output geographies, *International Journal of Geographical Information Science*, 12, 673-685

Openshaw, S. and Rao. L. (1995) Algorithms for reengineering 1991 Census geography *Environment and Planning A* 27 425-46

Rabe-Hesketh, S., Pickles, A. & Taylor, C (2000) Generalized linear latent and mixed models g129. *Stata Technical Bulletin*, 53, 47-57.

Rees P H (1994) Estimating and projecting the populations of urban communities. *Environment and Planning (Series A)*, 26: 1671-1697.

Rindskopf, D. (1992) A general-approach to categorical-data analysis with missing data, using generalized linear-models with composite links, *Psychometrika*, vol.57, no.1, pp.<sup>29-42</sup> Rubin, D.B. (1987) *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley and Sons

Robinson, G. (1950) 'Ecological correlation and the behavior of individuals', *American Sociological Review*, 15, 351-357

Schafer, J.L. (1995) *Analysis of Incomplete Multivariate Data by Simulation*. London: Chapman and Hall.

Simpson, S. (2000) Small area estimation using census data in the UK, in A. Dale, A., E. Fieldhouse and C. Holdsworth, *Analyzing Census Microdata*, pp.217-222, London: Arnold

Social Exclusion Unit (2000) *National Strategy for Neighbourhood Renewal: a framework for consultation*, SEU, April

Tranmer, M and Steel, D.G. (1998) 'Using census data to investigate local population structure'. *Environment and Planning (A)*, 30 817-83 1.

Ward, C. and Dale, A. (1992) Geographical variation in female labour force participation: an application of multilevel modelling, *Regional Studies*, Vol 26:3, pp 243-255

## Tables

**Table 1. Parameter estimates and sampling errors for different model specifications.**

	Logistic model with no geography	Multilevel, 30K threshold	Multilevel, 15K threshold	Multilevel, 7.5K threshold	Logistic + areas of 30k threshold
Indian	.45 (.20) *	.26 (.21)	.27 (.22)	.27 (.22)	.23 (.21)
Pakistani	1.31 (.14) *	1.09 (.15) *	1.04 (.15) *	1.06 (.15) *	1.08 (.15) *
Bangladeshi	.72 (.38)	.44 (.41)	.45 (.41)	.51 (.41)	.41 (.34)

**Table 2: Unemployment rates for economically active non-whites in Ambrosia SAM areas**

	M1	M2	M3	M4	M5	Truth
<b>Area 1</b>						
<b>Pred</b>	.00	.20	.12	.12	.11	.11
<b>Lo</b>	.00	.15	.09	.10	.09	
<b>Hi</b>	1.00	.27	.15	.14	.14	
<b>Area 2</b>						
<b>Pred</b>	.00	.19	.12	.13	.13	.12
<b>Lo</b>	.00	.15	.10	.11	.11	
<b>Hi</b>	1.00	.27	.16	.15	.16	
<b>Area 3</b>						
<b>Pred</b>	.00	.20	.11	.11	.11	.07
<b>Lo</b>	.00	.15	.09	.09	.09	
<b>Hi</b>	1.00	.27	.14	.13	.13	
<b>Area 4</b>						
<b>Pred</b>	.30	.30	.30	.28	.31	.33
<b>Lo</b>	.19	.22	.26	.26	.28	
<b>Hi</b>	.43	.38	.34	.31	.35	
<b>Area 5</b>						
<b>Pred</b>	.25	.23	.22	.22	.19	.23
<b>Lo</b>	.16	.17	.18	.18	.16	
<b>Hi</b>	.37	.30	.26	.26	.21	

M1 Logistic regression model of SAM data only

M2 Multilevel logistic model SAM data only

M3 Multilevel logistic model SAM + CAS margin for unemployment

M4 Logistic SAM + CAS margin for unemployment

M5 Multilevel logistic model SAM + CAS margins for unemployment and non-white

Truth True values of the unemployment rates for non-whites as obtained from 100% SAM data

**Pred** is predicted value, **Lo** is lower bound of 95% confidence interval, **Hi** is higher bound of 95% confidence interval.

**Table 3 Percentage of sample uniques which are also population unique by geography and sample size**

Key variables:

Age (19), sex (2), ethnic group (4), migration (4), economic status (5), tenure (6), marital status (5)

Sample	Area size				
	4K	8K	16K	32K	120K
2%	15.8	12.7	10.6	8.8	17.5*
3%	17.8	14.8	12.8	10.8	
4%	19.6	16.8	14.7	12.6	
5%	21.4	18.4	16.3	14.3	
10%	28.5	25.5	23.5	21.3	

Area size	4K	8K	16K	32K	120K
% population unique	10.49	6.90	4.42	2.77	8.96*

\* extended key: age (94), sex (2), ethnic group (10), migration (4), economic status (11), tenure (6), marital status (5)

	1991 SARs	SAM1	SAM2
Sample %	2	5	5
Threshold	120K	10K	5K
IND3a	0.056	0.068	0.136
IND3b	0.326	0.334	0.354
IND4a	0.043	0.087	0.096
IND4b	0.148	0.141	0.159
IND5	0.154	0.087	0.098
Mean	0.145	0.143	0.168

	1991 SARs	SAM1	SAM2
Sample %	2	5	5
Threshold	120K	10K	5K
IND3a	0.0074	0.0165	0.0223
IND3b	0.0187	0.0373	0.0440
IND4a	0.0067	0.0113	0.0174
IND4b	0.0135	0.0215	0.0281
IND5	0.0118	0.0099	0.0140
Mean	0.0116	0.0193	0.0252

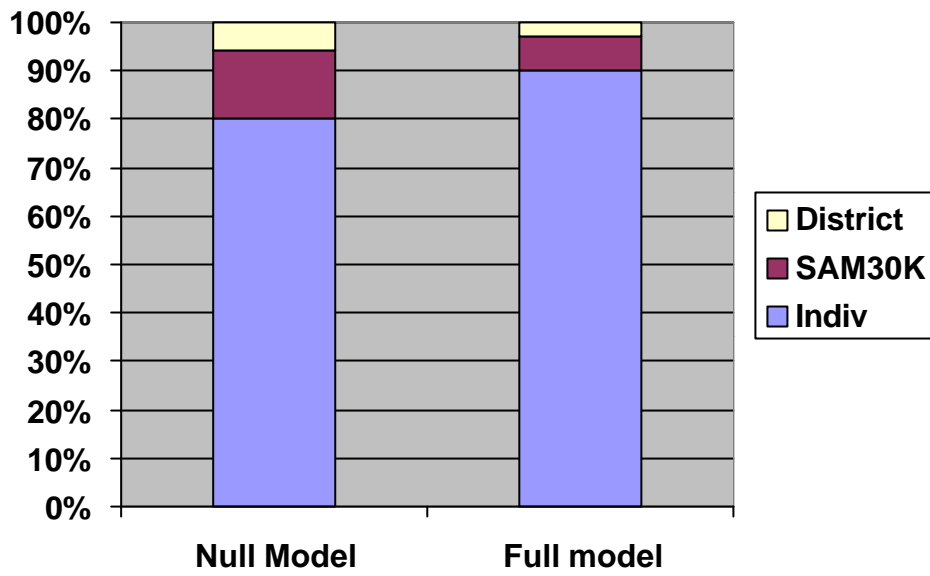
Table 6 Probability of a correct match given a unique match under several disclosure risk scenarios for 1991 SARs and two possible SAMs – CARNATION file			
	1991 SARs	SAM1	SAM2
Sample %	2	5	5
Threshold	120K	10K	5K
IND3a	0.0963	0.1616	0.1848
IND3b	0.3323	0.4660	0.4988
IND4a	0.0432	0.0868	0.1064
IND4b	0.1879	0.1555	0.1836
IND5	0.1407	0.1079	0.1265
Mean	0.1601	0.1955	0.2200

Table 7 Probability of a unique correct match under several disclosure risk scenarios for 1991 SARs and two possible SAMs – CARNATION file			
	1991 SARs	SAM1	SAM2
Sample %	2	5	5
Threshold	120K	10K	5K
IND3a	0.013	0.028	0.032
IND3b	0.019	0.046	0.047
IND4a	0.009	0.016	0.022
IND4b	0.016	0.029	0.033
IND5	0.015	0.017	0.022
Mean	0.014	0.027	0.031

Table 8 Relative risk of SAM1 and SAM2 compared to the SARs

Table	Metric	Analysis Frame	File	SARs	SAM1	SAM2
2	pr(cm um)	Additional Impact	Ambrosia	1.00	1.13	1.50
3	Pr(ucm)	Additional Impact	Ambrosia	1.00	0.52	1.06
4	pr(cm um)	Scenario	Ambrosia	1.00	0.99	1.16
5	Pr(ucm)	Scenario	Ambrosia	1.00	1.66	2.16
6	pr(cm um)	Additional Impact	Carnation	1.00	1.02	1.04
7	Pr(ucm)	Additional Impact	Carnation	1.00	0.42	0.87
8	pr(cm um)	Scenario	Carnation	1.00	1.22	1.37
9	Pr(ucm)	Scenario	Carnation	1.00	1.92	2.22
mean scenarios				1.00	1.45	1.73
mean additional impact				1.00	0.77	1.12
mean Carnation				1.00	1.14	1.38
mean Ambrosia				1.00	1.07	1.47
mean pr(cm um)				1.00	1.09	1.27
mean pr(ucm)				1.00	1.13	1.58
Mean all				1.00	1.11	1.42

Figure 1 Variance components for three level model of unemployment for 7 test areas



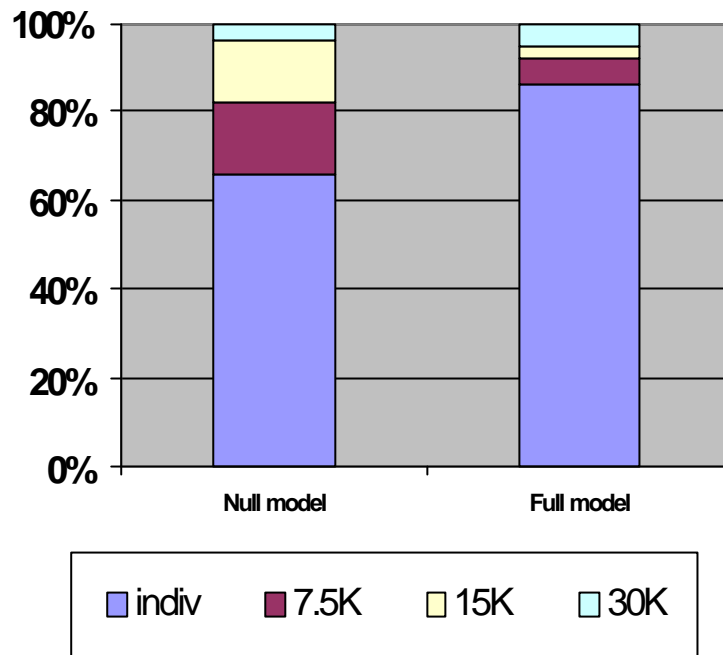
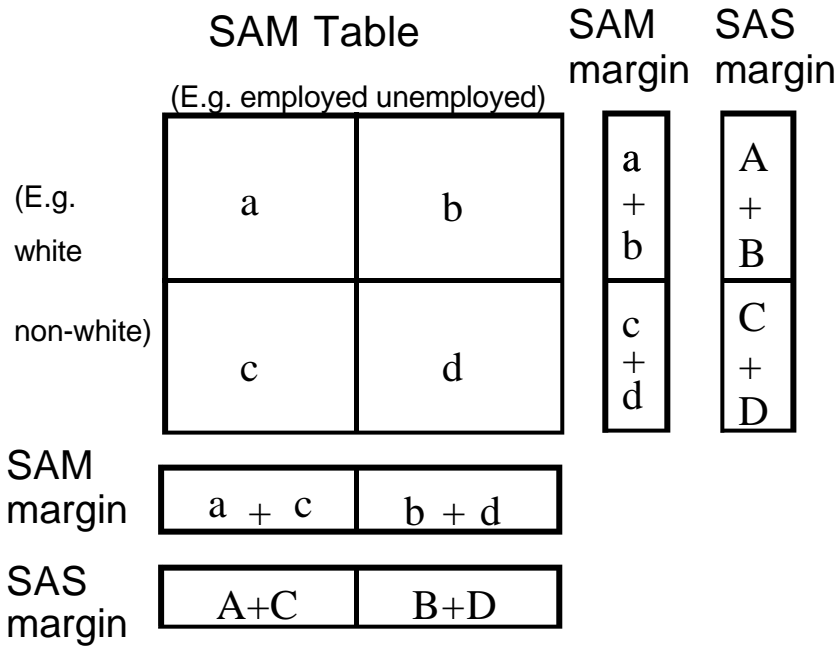


Figure 2. Decomposition of variance in a four level model of unemployment for 'Ambrosia'

## Appendix 1

Figure 1: schematic diagram, showing how SAM and CAS data may be combined to estimate quantities of interest for particular SAM areas.



## Appendix 2: The Disclosure Intrusion Simulation (DIS) method

The basic principle of the DIS method is to remove a small number records from the target microdata file and then copy back some of those records, with each record having a probability of being copied back equal to the sampling fraction of the original microdata file. This creates two files, a new slightly truncated target file and a file of the removed records which is then matched against the target file. The method has two computational forms, the *special form*, where the sampling is actually done and the *general form*, where the sampling is not actual done but the equivalent effect is derived using the partition structure of the microdata file and sampling fraction.

### *The special method*

The special method follows the following five-step procedure, (a schematic version can be found in Appendix B).

- (i) Take a sample microdata file (A) with sampling fraction  $S$ .
- (ii) Remove a small random number of records (B) from A, to make a new file (A').
- (iii) Copy back a random number of the records in B to A' with each record having a probability of being copied back equal to  $S$ .

The result of this procedure is that B will now represent a fragment of an outside database (an identification file) with an overlap with the A' equivalent to that between the microdata file and an arbitrary identification file with zero data divergence (with no differing values for the same individual).

- (iv) Match B against A'. Generate an estimate of the matching metrics particularly, the probability of a correct match given a unique match,  $pr(cm|um)$ , between the fragment.
- (v) Iterate through stages i-iv until the estimate stabilises.

### *The general method*

A more general method can be derived from the above procedure. Imagine that the removed fragment (B) is just a single record. There are six possible outcomes depending on whether the record is copied back or not and whether it was a unique, in a pair or in a larger partition class.

*Table 1: Possible per record outcomes from the DIS general method*

record is:	<i>Copied back</i>	<i>not copied back</i>
<i>sample unique</i>	<b>correct unique match</b>	non-match
<i>one of a sample pair</i>	multiple match including correct	<b>false unique match</b>
<i>one of a larger equivalence class</i>	multiple match including correct	false multiple match

The critical cells in the above table are those where a unique record is copied back and where one of a sample pair is not. The relative numbers in the cells determine the probability of a correct match given a unique match;  $pr(cm|um)$ . Given this, it is possible to shortcut the special method since one can derive an estimated probability of a correct match given a unique match from:

$$pr(cm|um) \cong \frac{U * f}{U * f + P * (1 - f)}$$

Where U is the number of sample uniques, P is the number of records in pairs and  $f$  is the sampling fraction.



### Appendix 3 Proposed SAM specification

Dataset	1991 SARs	SAMSPEC	Comment
<b>Sample %</b>	<b>2</b>	<b>5</b>	
<b>Threshold</b>	<b>120K</b>	<b>5K &amp; 10K</b>	
Age	94	19	5 year age-groups
Type of community establishment	14	-	
Status in Community establishment	3	2 (or omit)	Communal/non communal
Country of birth	42	3	UK/EU/Non EU
Distance of travel to work	9	3	Classification used in 2001 tables
Distance of move (migrants)	14	3	Classification used in 2001 tables
Economic Position(primary)	10	10	Classification used in 2001 tables
Economic Position(secondary)	8	-	Omit
Ethnic Group	10	5	Classification used in 2001 tables
Family type	8	4	
Gaelic language	5	-	Omit
Usual hours of work	73	2	Full-time/part-time
Industrial Classification	61	-	Omit
Long term Limiting Illness	2	2	
Marital Status	5	3	Single/partnered/previously married
Area of former Residence	13	3	
Occupational Classification	73	-	Omit
Number of highest Qualifications	3	-	Omit
Level of highest qualification	3	2	Will include extra detail on new question
Subject of highest qualification	35	-	Omit
Relationship to household head	8	2	HOH indicator (0/1)
Resident Status	3	-	Restrict sample to usual residents
Sex	2	2	
Social Class	9	9	this should count as 3 or 4 categories for confidentiality.
SEG group	20	-	Omit
Term-time address	4	-	Omit
Tranwork	10	5	Car/public/bike/foot/other
Welsh Language	5	-	Omit
Work Place	5	-	Omit
Bath/Shower	3	2	Yes/no
Central Heating	3	2	Yes/no
Inside WC	3	-	Dropped in 2001
Number of Cars	4	3	0/1/2+
Household Dwelling Space type	14	7	
Number of Residents per room	5	3	<1/1-1.5/>1.5
Tenure of household space	10	5	Own/mortgage/la/HA/private rent
Number of Residents	4	4	
Number of dependant children	2	2	No dep. Children/dep. children present
No with LTILL	2	2	Calculate mean WITH AND WITHOUT VARIABLE
No of residents of pensionable age	2	2	Calculate mean WITH AND WITHOUT VARIABLE
No of residents in employment	3	2	Calculate mean WITH AND WITHOUT VARIABLE
Economic position of family head	3	-	

Sex of family head	2	-	
Social Class of family Head	9	-	

### Appendix 3      Variation in ward size

The distribution of the population by ward size can be considered in two useful ways

- a) the proportion of the **population** living in wards (postcode sectors in Scotland) that exceed the given population threshold (5, 10 and 15k)
- b) the proportion of **wards** (postcode sectors) with a population size that exceeds the given threshold.

#### Key points :

##### for England and Wales

- 70% of the population live in wards greater than 5K (42% of all wards)
- 32% of the population live in wards greater than 10K (13% of all wards)
- 9% of the population live in wards greater than 15K (2% of all wards)

##### for metropolitan areas only (West Yorkshire, South Yorkshire, Merseyside, Gt Manchester, West Midlands, Tyne and Wear, Outer London, Inner London)

- 99% of the population live in wards greater than 5K (96% of all wards)
- 70% of the population live in wards greater than 10K (56% of all wards)
- 23% of the population live in wards greater than 15K (13% of all wards)

This masks substantial differences between separate metropolitan areas e.g. the mean population size of wards in West Yorkshire (15.8K) is almost twice that for Inner and Outer London (8.1 and 8.9K)

##### for Scotland

- 81% of the population live in postcode sectors greater than 5K (50% of all postcode sectors)
- 22% of the population live in postcode sectors greater than 10K (9% of all postcode sectors)
- 3% of the population live in postcode sectors greater than 15K (1% of all postcode sectors)